METHOD FOR SCANNING, ANALYZING AND HANDLING VARIOUS KINDS OF DIGITAL INFORMATION CONTENT

Abstract of the Disclosure

Computer-implemented methods are described for, first, characterizing a specific category of information content—pornography, for example—and then accurately identifying instances of that category of content within a real-time media stream, such as a web page, e-mail or other digital dataset. This content-recognition technology enables a new class of highly scalable applications to manage such content, including filtering, classifying, prioritizing, tracking, etc. An illustrative application of the invention is a software product for use in conjunction with web-browser client software for screening access to web pages that contain pornography or other potentially harmful or offensive content. A target attribute set of regular expression, such as natural language words and/or phrases, is formed by statistical analysis of a number of samples of datasets characterized as "containing," and another set of samples characterized as "not containing," the selected category of information content. This list of expressions is refined by applying correlation analysis to the samples or "training data." Neural-network feed-forward techniques are then applied, again using a substantial training dataset, for adaptively assigning relative weights to each of the expressions in the target attribute set, thereby forming an awaited list that is highly predictive of the information content category of interest.